

Frequently asked questions (FAQ)

1. Where and how to download website HTML files?

You can download them from <https://drive.google.com/drive/folders/1ZPWjQZMB7xeegjaeZfzqeJKq9oKuukio>. Please use your USC email to request access to your news website files. There are files for only 3 news sites Newsday, NBC_News and FOX_News for this assignment. C-Span and Chicago Tribune had far fewer files as the first 3 sites and so to have equal load for everyone, they are not given. You should download the folder for the news site that is assigned to you based on your USC ID.

USC ID ends with	News Sites to Index
01-40	Newsday
41-70	FOX_News
71-100	NBC_News

2. The instruction of hw4 suggests that we should install ubuntu on windows so as to install solr more conveniently. However, if I do not want to install ubuntu, may I install solr and other software needed directly to windows system?

You can. However, if you run into some issues when installing directly on Windows, we may not be able to guide you. You will be on your own.

3. Are we allowed using any language and platform instead of PHP for the Web pages?

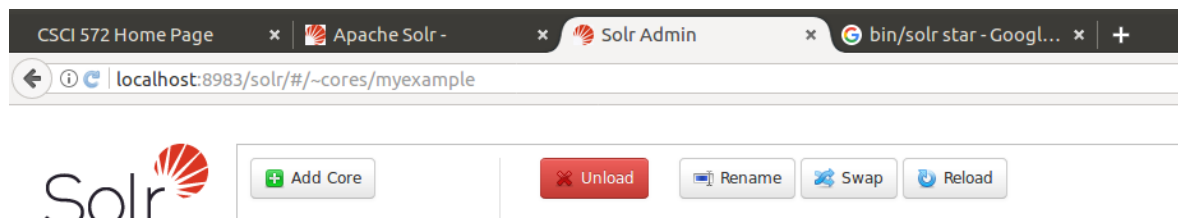
Yes, you can use node.js or any other language, but you need to check its compatibility with Solr at your end only. However no help will be provided for the code implementation for this homework.

4. There is no step-wise instruction given for installing solr. Can you guide me in installation process? I downloaded Solr-7.1.0 zip.

You need to unzip it and then follow the instructions in "Using Solr to Index a Web Site" document. Once you are done with the Ubuntu installation steps, the Solr setup instructions are to be executed inside Ubuntu. So, setting up Java 8, downloading, unzipping Solr, Apache, PHP etc. all should be done inside Ubuntu.

5. How to reload our core and query with new configuration.

There will be a reload button in the Solr instance's core admin: Explore the Solr instance, it's various configurations and settings.



6. Do I need to install an Apache server on our computer in addition to running the Solr server?

Apache server will be installed by default in Linux machines. In the assignment we are using PHP script to communicate to the Solr. As we know PHP is a server-side scripting language and it requires Apache or any other web servers.

However, if you run into any problem then use the below commands on Ubuntu terminal.

To install Apache on Ubuntu

```
sudo apt-get install apache2
```

To install PHP on Ubuntu

```
sudo apt-get install libapache2-mod-php
```

7. Can I run solr in cloud mode?

You don't need cloud mode for this assignment. Only use "bin/solr start" to start running the Solr instance in stand-alone

To Restart Solr-

First do, "bin/solr stop -all". Then "bin/solr start"

8. I am not mentioning my top 10 results in the report. I am maintaining separate text files for each query. Is that allowed or shall I mention the results only in the report?

It would make grading much easier if you included all in one file.

9. I am getting zero overlaps between Default and pagerank algorithms. Could you please let us know how many points would be deducted for zero overlaps?

As long as you follow the assignment steps and implement it correctly, you won't be penalized. Just make sure your steps are clear in your report and you should be fine.

10. How to use additional parameters from code?

Here is an example of using Additional Parameters:

Please notice the below parameters when querying Solr from the dashboard: q, fq, sort, etc.,

Solr

- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- myexample1
- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema
- Segments info

Request-Handler (qt)

/select

common

q

Harry Potter

fq

sort

score asc

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

indent

debugQuery

When querying Solr from the dashboard, you enter the query in the q parameter input box-> "Harry Potter".

You may use the other parameters as well to further refine the query.

Ex: sort parameter's default value when nothing is specified is "score desc", i.e. the results are sorted and displayed from highest score to lowest score. You can change to "score asc".(For demo purposes)

How do you specify this in the PHP code? By sending it as an additional parameter in the call for Solr:

```
$additionalParameters = array(
'sort' => 'score asc'
);
$results = $solr->search($query, 0, $limit, $additionalParameters);
```

URL

->

http://localhost:8983/solr/myexample1/select?indent=on&q=Harry Potter&sort=score asc&wt=json

Hope this clears your doubts on how to use additional parameters from code, and why you were getting the error.

Further reading -> <https://wiki.apache.org/solr/CommonQueryParameters>

11. Do we need sort results by any attribute in the first algorithm?

No. It is already sorted by default value -> "score desc".

12. How to use NetworkX and JSOUP?

Both libraries need to be used in different parts of the assignment. JSOUP is used in JAVA program to extract links from given HTML pages and create edge list file. NetworkX is used in Python to calculate pagerank for all pages.

13. What is the use of Java code?

The Java code creates the edgelist.txt file. You must use this as input for Python code, which creates a graph using NetworkX library.

14. Do we have to write page rank algorithm?

The default search result is returned based on Lucene's ranking algorithm.

The next step is to get the results based on the page ranking score. You do not need to write the algorithm yourself. It is implemented in NetworkX library. You use it to obtain an external file with page ranking scorings, which you add to Solr. Please see the document.

15. If that page doesn't contain description information, we don't need to show description?

That's ok. You can show Null or N/A.

16. Getting much better results with 'pageRankFile asc' as opposed to 'pageRankFile desc'. Which one shall I use?

Solr uses Boolean model to narrow down the range of the documents that needs to be scored. This will be done based on the query term. Page rank "desc" will next sort and return these from pages from highest score to lowest score.

The pages having highest page rank scores (maybe a homepage, or some other popular page with many links to it) may not be most relevant to query term.

17. What should be the format of score in external_pageRankFile?

It can be exponential or decimal format, but exponential is better as it helps Solr(Lucene) in ranking the pages properly.

18. Handling duplicates?

Filter duplicates based on the ID of the returned page. If you find duplicate URLs in top 10 results, you can just handle them as if they are unique (No need to worry about it during calculations or display). Sometimes news websites publish the same page at 2 different places in their domain and forget to take old one down after the new one is published.

19. How should the front end look like?

You are free to implement any UI which resembles the results returned by any search engine.

20. What all information should be shown after search button is pressed?

Title, URL(og_url), ID, Description(og_desc), If Description is not available, just display NA. If URL is missing, for HW4, fetch from the provided map file. Title and URL should be clickable.

21. Should we use quotes with 8 queries?

No, don't use any quotes while entering the queries.

22. Are we allowed to use whichever Solr Client Libraries within PHP and Node.js categories on this page <https://wiki.apache.org/solr/IntegratingSolr>?

Yes, you can use any as long as search works fine.

23. what should be format in pageRankFile.txt file?

`<full_file_path/doc_id>=<page_rank_score>`

24. Where can I find more information about how “href:abs” works?

<https://jsoup.org/cookbook/extracting-data/working-with-urls>

25. What should be the naming convention for page_rank file?

It should be external_<filename>.txt also your pagerank filename and field name in managed Schema should be same.

26. Where can I check if the Solr results doesn't show up properly?

Check error logs in `/var/log/apache2/error.log`

27. Order of the search results in the display page

Order doesn't matter but should contain the following results

1. Title (clickable)

2. URL (clickable)

3. ID

4. Description

28. Should we consider src and link[href] for edgeList creation?

Consider only `a[href]` as it is in the example given in assignment

29. external_pageRankFile should be the same for everyone for a given News website?

Yes

30. deliverables for home work 4

1. The external page rank file

2. The report containing all the written details - screenshots, table etc.

3. The working source code in a folder.

DO NOT include the code you used to generate EdgeList and PageRank values

